DOCUMENT RESUME

ED 333 747 FL 019 246

AUTHOR Griffin, Patrick
TITLE Characteristics of the Test Components of the IELTS Battery: Australian Trial Data.
PUB DATE Apr 90
NOTE 17p.; Paper presented at the Regional English Language Centre Seminar on Language Testing and Language Program Evaluation (Singapore, April 9-12, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *English (Second Language); Foreign Countries; Grammar; Interrater Reliability; *Language Tests; Listening Skills; Reading Tests; Speech Skills; *Test Construction; Testing; *Test Reliability; *Test Validity; Vocabulary; Writing Tests
IDENTIFIERS *Australia; *International English Language Testing System

ABSTRACT

Results of the International English Language Testing System (IELTS) battery trials in Australia are reported. The IELTS tests of productive language skills use direct assessment strategies and subjective scoring according to detailed guidelines. The receptive skills tests use indirect assessment strategies and clerical scoring procedures. Component tests in reading, writing, listening, speaking, and grammar and vocabulary were developed by international teams for use in measuring English language competence and identifying suitable candidates for study in English-language-medium programs. The report describes the trial subject sample and test component characteristics, and presents and discusses detailed statistical results for each test item, reliability statistics, and data on inter-test correlations and interrater reliability. The grammar and vocabulary component was removed from the test, and some item deletions are noted. A brief list of references is supplied. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED333747

Characteristics of the Test Components of the IELTS Battery: Australian Trial Data.

Patrick Griffin

BEST COPY AVAILABLE

Paper presented at the Regional English Language Centre, Annual Seminar, Singapore, April 9-12 1990.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Griffin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FL019246

The Nature of The Test Battery

In a paper on the IELTS development presented at the fifth ALAA conference in Launceston, the structure, nature and procedures adopted in developing and trialing the test components of the International English Language Testing System battery were described (Griffin, 1988). This paper, addresses the results of the trials of the test battery using the data collected in the Australian component of the trials. The testing system focuses on both productive and receptive skills. The tests of productive skills employ direct assessment strategies and use subjective marking procedures guided by detailed guidelines and training of assessors. The tests of receptive skills employ indirect assessment strategies and employ clerical marking procedures. Conversion to final band scores is based on the judgements of the test developers which was informed by knowledge of the candidates and the skills assessed by the tasks set and the directions given to test item writers. Several workshops were used to develop training methods, criteria and rating protocols for the productive skills of speaking and writing. Assessments of these skills were interpreted as being at one of ten levels or bands as described in the specifications of the tests. These were labelled from band 0 to band 9. Band 0 indicated no proficiency or a failure to take the test and band 9 indicated the highest level of language proficiency, roughly equivalent to a native- like proficiency. This did not presume however that native speakers would always score at the highest levels.

Direct interpretation of the receptive skills were not possible. Indirect assessment, based on paper and pencil tests were used. The total test scores were then used as necessary information to estimate the band level of these skills. Definitions of the band levels were included in the specifications of the tests. The reading and writing tests were designed with specific academic populations in mind. A series of specifications for special purpose modules focussed on sub populations in academic fields including Science and technology, Art and Social Sciences and Life and Medical Sciences. A further set of specifications was developed to cater for what was described as a non academic, general training population. The reading and writing tests for each special population were contained within the same test booklet but have been ordered such that all reading tasks are completed before writing tasks could be attempted.

The component tests were developed from specifications written by teams from Australia, Canada and the United Kingdom. The battery of tests were designed to measure English Language Competence and to identify suitable candidates for study in programs conducted in an English language medium. Five tests were originally included in the battery of tests which an individual candidate could expect to take. These were:-

1. Reading
2. Writing
3 Listening
4. Speaking
5. Grammar and Lexis.

The fifth test, that of grammar and lexis, has now been omitted from the test battery.

Table 1
IELTS Battery Composition.

Component	Code	Population Focus
Grammar and Lexis	G1	General
Listening	G2	General
Speaking	G3	General
Reading	M1	Science and Technology
Reading	M2	Arts and Social Sciences
Reading	M3	Life and Medical Sciences
Reading	M4	General Training
Writing	M1	Science and Technology
Writing	M2	Arts and Social Sciences
Writing	M3	Life and Medical Sciences
Writing	M4	General Training

The tests were administered throughout Australia and South East Asia by Australia's International development Program of the Universities and Colleges (IDP). British Council representatives also trialed the test in non English speaking countries. The overall coordination of the trials of the test was conducted at the University of Lancaster by the IELTS project team. This report focuses on the data gathered by the Australian contribution to the trial forms of the IELTS. The schedule of the IELTS trials are presented in the following table.

Table 2
The Schedule of Testing in the IELTS Trials.

Code	Component	Items	Time (mins)
G1	Grammar and Lexis	38	30
G2	Listening	41	30
G3	Speaking	n/a	15
M1	Reading	38	50
M1	Writing	2	40
M2	Reading	39	50
M3	Writing	2	40
M3	Reading	33	50
M3	Writing	2	40
M4	Reading	42	50
M4	Writing	3	40

All tests were group administered except the test of Speaking. This was of an interview format and was individually administered. The schedule kept the total testing time at 110 minutes and allowed the full group testing battery to be administered in one sitting. Not all candidates in the trials were asked to complete the full battery. The purpose of the trials was to establish the properties of the components and to establish a basis for future reliability and validity studies.

The Trial Samples

Trial testing, under the direction of the Australian office of the IDP took place in four countries- Indonesia, Thailand, Hong Kong and Australia. In Hong Kong and Australia, native speakers were assessed. Table 3 presents the number of candidates assessed on each test in each of the countries from which samples were drawn.

Table 3
Sample Sizes for Each Component Test of IELTS and Place of Administration.

Country	Test							Total
	G1	G2	M1	M2	M3	M4	M5	
Hong Kong	482	465	261	105	113	121	0	1547
Indonesia	105	106	77	67	73	69	0	597
Thailand	45	47	8	10	8	21	0	139
Australia	201	131	270	257	283	381	124	1647
Total	843	749	616	439	477	592	124	3930

Test Characteristics: General

A difficulty presents itself in a presentation of results about the development and trialing of the IELTS. Because of the security of the test, it is not possible to illustrate data using examples of test items. The data on each test was analysed to provide, item and total means, reliability and point biserial correlation coefficients for each item. Candidates were also asked to rate themselves on a nine point scale to gain a self assessment estimate of their band scale. This estimate is presented in the Table as SELF. In each test some additional questions were asked of the students. These were used for feedback to the test developers and the means, standard deviations and correlations with the test total score are also reported in these analyses. The questions were.

FB1 Do you feel that this was a fair test of your English?
FB2 Was there enough time for you to complete the test?
FB3 Was the test too hard?
FB4 Was the test too easy?
FB5 Were the questions realistic?
FB6 Were the instructions clear?

Item FB5 was not asked for the Grammar and Lexis test. Two tables and a figure are presented for each test in the IELTS battery. The first Table presents the following information for the General Training Module. This paper presents the results of the analysis of this module. Other test module results will become available as the manuals are released by the management of the IELTS project and general data fro the modules based on the Australian data were presented by Griffin (1989). The general results will encompass both the UK data and the Australian data and may not be identical to the results presented in this paper. Large differences would not be expected however. The table below p[resents the general characteristics for the IELTS trials without presenting the specific item level data.

Table 4
General Characteristics of Modules in IELTS

Module	N	Items	Mean	SD	Alpha	p		phi		Rasch diff		items≤6
						max	min	max	min	max	min	
G1	843	38	26	6.4	82	979	230	626	114	-3.31	2.73	1368
G2	749	41	23.7	7,5	83	955	116	628	044	-2.78	2.88	1270
ASS	616	38	17.3	8.9	90	787	116	654	204	-1.83	2.13	1950
LMS	439	39	15.8	9.4	92	758	075	690	287	-3.06	2.47	1853
ST	477	33	14.9	7.9	90	790	073	686	307	-1.36	2.96	1458
GT	592	42	25.2	6.7	80	934	212	547	145	-2.15	2.01	880

The above data illustrate the consistency across modules. They are of uniformly high reliability, have a wide range of item difficulty and discrimination and have suitable levels of fit to an underlying dimension

as estimated by the proportion of item which fit the Rasch model. In addition to the test level data above, specific item level data was collected on the feedback items.

(i) The feedback from the candidates regarding the suitability of the test for their purposes and the candidates' perception of the fairness of content, time available, clarity of instructions and ease or difficulty of the instrument. Where both reading and writing are presented, the same items are asked for each skill. The feedback items were based on a dichotomous response scored '1' for 'yes' and '0' for 'No'. So the higher the value, the greater satisfaction of the candidate.

(ii) Estimates of internal consistency coefficients of reliability (alpha), the number of cases providing data for the test and the overall average score on the test.

(iii) Standard deviations and point biserial correlations for each item are also presented.

Table 5
General Training Reading and Writing Test:
General Properties and Student Feedback

Variable	Mean	SD	r.tot
M4RFB1	1.262	.440	-.045
M4RFB2	1.658	.474	-.374
M4RFB3	1.552	.497	.286
M4RFB4	1.970	.170	.030
M4RFB5	1.157	.364	-.008
M4RFB6	1.131	.338	-.166
M4RSELF	4.587	1.450	.244
M4W1	4.293	1.184	.342
M4W2	4.453	.891	.369
M4W3	4.256	1.037	.301
M4WFB1	1.237	.425	-.069
M4WFB2	1.575	.495	-.264
M4WFB3	1.554	.497	.183
M4WFB4	1.963	.188	.031
M4WFB5	1.124	.330	-.114
M4WFB6	1.100	.300	-.178
M4WSELF	4.025	1.357	.211
M4TOT	25.182	6.715	
ALPHA	.845		
N OF CASES	592		

The second table provides information on each test item. The data provided are the item mean, standard deviation and the point biserial coefficient.

Table 6
General Training Test of Reading:

	MEAN	S.D.	r.pbi	LOGIT	ERROR	FIT
M4A2	.859	.347	.306	-1.21	.13	.35
M4A3	.917	.275	.184	-1.91	.16	-.04
M4A5	.848	.359	.270	-1.18	.13	.03
M4A6	.800	.399	.291	-0.80	.11	.09
M4A7	.473	.499	.337	0.81	.09	.45
M4A8	.886	.317	.240	-1.51	.14	.05
M4A9	.861	.345	.403	-1.62	.16	-.64
M4A11	.853	.354	.343	-1.17	.13	-.59
M4A12	.658	.474	.370	.07	.10	-.70
M4A13	.304	.460	.357	1.73	.10	-.91
M4A14	.604	.489	.470	.26	.10	-1.86
M4A15	.888	.315	.339	-1.56	.15	-.40
M4A16	.366	.482	.364	1.44	.09	-.99
M4A17	.934	.248	.321	-2.15	.19	-1.12
M4A18	.922	.267	.267	-1.93	.17	-.57
M4A19	.903	.295	.444	-1.87	.18	-1.44
M4A20	.864	.342	.370	-1.25	.19	-.72
M4A21	.841	.365	.356	-1.06	.12	-.66
M4A22	.636	.481	.293	.15	.10	1.34
M4A23	.814	.389	.304	-0.93	.12	.04
M4A24	**.542**	**.498**	**.145**	**.63**	**.10**	**6.12**
M4A25	**.511**	**.500**	**.485**	**.70**	**.10**	**-4.45**
M4A26	.574	.494	.361	.38	.09	-.23
M4A27	.768	.422	.431	-.74	.12	-.32
M4A28	.613	.487	.480	.11	.10	-1.55
M4A29	**.432**	**.495**	**.323**	**.99**	**.10**	**2.36**
M4A30	**.488**	**.500**	**.314**	**.67**	**.10**	**3.16**
M4A31	**.241**	**.428**	**.212**	**1.99**	**.11**	**2.50**
M4A32	.290	.454	.285	1.70	.10	1.29
M4A33	.694	.461	.465	-1.03	.14	-.82
M4A35	.278	.448	.412	1.68	.10	-1.08
M4A36	**.356**	**.479**	**.533**	**1.24**	**.10**	**-3.67**
M4A37	**.310**	**.463**	**.205**	**1.52**	**.11**	**4.26**
M4A38	.212	.409	.274	2.01	.11	1.25
M4A39	.584	.493	.540	-.01	.11	-1.73
M4A40	.572	.495	.493	-.06	.11	.20
M4A41	.295	.456	.426	1.48	.10	-1.03
M4A42	.456	.498	.547	.43	.10	-1.95
M4A43	.234	.424	.344	1.70	.11	1.22
M4A44	.413	.492	.486	.67	.11	-.19
M4A45	.469	.499	.533	.21	.11	-.98
M4A46	.599	.490	.495	-.71	.18	.00
Mean	24.68	6.73				
Alpha	0.79					

The general training module has a wide range of difficulty. From the table and the figure, it is evident that the test caters for the suitable range of candidates and discriminates at the appropriate levels. Not all items fit the latent trait scale. Seven of the 42 items do not appear to be measuring the same dimension of language as the other items. However, the remaining 35 items are, according to their fit to the Latent

trait, acting together to assess language ability of the candidates. This is despite the fact that the candidate group was obtained from a wide range of backgrounds, first languages and prospective courses. The test appears to have sound construct validity. In earlier studies of reading tests using Item response theory as a guide to construct validity Andrich and Godfrey (1978) analysed Davis' test of reading comprehension. Their analysis argued that 80 percent of the items fitting the underlying trait gave sufficient evidence of construct validity. In this case, the percentage is 83.3 percent. Hence the majority of items in the test are measuring the same construct. Construct validity would appear to have been demonstrated. The items which do not fit the underlying trait were also examined. Each involves the elimination of negative options of the elimination of distracting information. The block of items which contained most of these difficulties was eliminated from the final form of the test.

The test was clearly not difficult overall. Apart from one set of items, M4A31 to M4A38 the items have high mean scores. The more difficult items have now been removed from the test as well, largely because of the types of tasks used in the items. Hence the overall difficulty of the test has been reduced somewhat after the trials and the expected mean scores will rise.

The Figure below illustrates the distribution of the scores of the students relative to the distribution of the difficulty levels of the items on the test. Where the student distribution appears to be above the item distribution, it appears that the test may be too easy for the candidates as a whole group. This information needs to be taken into account when interpreting the feedback item information.

There are three scales in the figure. The first is the raw score of the students. The second is the latent trait logit scale and the third is the band scale for interpretation of the IELTS. This scale is an interval scale, based on the interval properties of the latent trait and is a linear transformation of the latent trait logit scale. The logit scale is derived from the application of the simple logistic model of the Rasch latent trait theory. It is computed from the equation

$$P_{(x|\beta_v,\delta_i)} = \frac{e^{(\beta_v-\delta_i)}}{1+e^{(\beta_v-\delta_i)}}$$

Knowledge of the characteristics of the student groups and identification of native speakers and their test performances were used to establish these levels. Like the assessment of the productive skills of speaking and writing, a professional judgement is ultimately required to transform the raw test scores of the receptive skills onto the band scales used for reporting to consumers of IELTS information.

Figure 1
General Training Test of Reading: Conversion from Raw Score to Band Levels.

RAW	LOGITS	PERSONS-	-ITEMS	BAND
	5.0			
41		X		
	4.0			
40		X		
39		X		
	3.0			
38		X		6
		XX		
37		XX		
36		XXX		5.5
35	2.0	XXXXXX	XX	
33 34		XXX	XXX	
32		XXXXX	XX	5
31		XXXX	XX	
30		XXXXXXX	X	4.5
29	1.0	XXXXXXXXXX	XX	
28 27		XXXXXXXX		4
26		XXXXXXXX	XXXX	
25 24		XXXXXXXX	XX	
23 22		XXXX	XXXX	3.5
21	.0	XXXX	XXX	
20 19		XXX		3
18		XX		
17 16		XXXXX		
15		X	XXX	2.5
14	-1.0	XXX	XXX	
13 12		X	XXXX	
11		X		
10		XX	XXX	2
9		X	X	
8 -2.0		X	XX	
			X	

Correlations among the different modules of the IELTS were all obtained as were correlations of the IELTS battery tests with other criterion measures. Existing records were used to obtain scores from the Hong Kong Examinations Authority for their listening test, the overall GCE grade in English, a summary score, comprehension score and a compositional score. This enables correlations to be obtained against all other scores. Where available, scores on the TOEFL, the Short Selection Test (SST) the ASLPR (ASLR AND ASLW for reading and writing estimates), the existing ELTS and the Oxford tests forms 2 and 3 forms A and B were obtained (O2A, O2B,O3A O3B). Self assessment was also gathered in that the students were asked to place themselves on a 9 point scale, but without any guidance as to the meaning of levels. These are labelled as SPR and SPW for self proficiency in Reading and Writing. Nevertheless, these scores enabled further insight into the behaviour of the IELTS battery against a range of other measures. Table 6 below presents the correlations of the IELTS battery with the criterion measures. Most of the emphasis is placed on the general training module as with the rest of the paper, and other criterion correlations will be made available as the manuals and other papers become available from the IELTS management. No correlations between the speaking test and other measures were obtained during the Australian trials.

CORRELATIONS IELTS 1989

	G2	M1R	M1W	M2R	M2W	M3R	M3W	M4R	M4W	
M4R	772	588		593		430				M4R
N	242	71		68		48				N
M4W	475					256	577	449		M4W
N	201					16	7	222		N
ELTS	826	258	388	448	712	203	273	524	446	ELTS
N	11	23	22	12	12	11	9	6	6	N
TOEFL	804	879	678	704	569	866	619	647	702	TOEFL
N	66	15	16	18	19	21	21	6	6	N
SST	-753			-269	-536		-760	-696	-715	SST
N	39			24	24		9	27	21	N
O2A	492									O2A
N	136									N
O3A	510									O3A
N	54									N
HKGRADE		-602	-614	-446	-460	-416	-411	-297	-504	-216
HKGRADE										
N	218	60	60	48	48	62	62	30	29	N
HKSUMRY			638	402	441	507	419	314	358	0
HKSUMRY										
N		60	60	48	48	63	63	30	29	N
HKLIST		484								
HKLISTN										
N	218									N
HKCOMPOS			531	407	248	372	117	282	464	0
HKCOMPOS										
N		60	60	68	48	63	63	30	29	N
SPR	406	404	508	472	562	363	384	254	192	SPR
N	402	225	231	98	87	104	104	342	177	N
SPW		351	460	475	520			149	235	SPW
N		189	189	94	93			219	145	N

While many of the sample sizes are small, the correlations are encouraging. Moderately high and appropriately signed correlations have been obtained with all modules with the TOEFL, the SST, the Oxford tests and the Hong Kong GCE Examination results. Too few cases were obtained to make any interpretation of the ASLPR ratings. This however should be easy for the IELTS Australia to remedy in the future. The evidence is encouraging for the IELTS battery in terms of criterion validity. It is clear that the IELTS is measuring language proficiency in the same domain measured by similar test batteries.

The correlations between te reading tests in the modules are also generally high, indicating that the tests are generally measuring the same underlying variable. This has been further explored by Alderson (1990) in his comparison of the Australian data with the combined UK and Australian data. The intercorrelations among the reading modules are presented in Table 8 below.

Table 8
Intercorrelations Among the Reading Tests.

	Arts	Sci	Gen Trng	Gram	List
Life Med	58	65	59	58	66
	(90)	(114)	(68)	(88)	(88)
Arts		47	58	69	62
		(100)	(74)	(198)	(198)
Sci Tech			49	80	79
			(60)	(123)	(123)
Gen Trng				78	77
					(123)
Gramm				79	
					(123)

Two things are noticeable. First , the generally low correlations of the general training module with the other reading tests and second, the generally high correlations of the grammar test with all other tests. Alderson, also illustrates this relationship and classifies the grammar test as a reading test, as is the listening test.

Reliability:

Reliability can be assessed from two aspects. First there is the classical internal consistency reliability estimates, and second there are the item level reliability or error estimates available from the latent trait analyses. Table 5 presents the error estimates and the internal consistency estimate of 0.79. The latent trait analyses illustrates the high item level reliability given that few item exceed errors of 0.20. These figures illustrate the reliability of the reading tests in the IELTS battery and in particular the reliability of the General training module. Reliability estimates assisted in the decision to remove the grammar and lexis test from the test battery.

The test of lexis and grammar was omitted from the IELTS battery after examination of reliabilities and after examination of issues underpinning the test. The four remaining tests all assess either a productive or receptive language skill. The test of grammar and lexis tested knowledge about language rather than the ability to use it for communicative purposes. In addition, there was no suitable scale of progression which could be developed for interpretation and reporting as with the other tests. While professional judgement is ultimately needed for reporting the levels of attainment on the reading and listening tests in terms of IELTS band levels, no similar translation could be provided for the test of lexis and grammar. These substantive reasons together with the lack of contribution to reliability beyond that which could be achieved by increasing the number of items in the reading test. This helped the management of IELTS to decide to recommend its omission from the battery. The table below illustrates the contribution of the lexis and grammar test to the overall battery of clerically scored tests and the overall reliability of the combined tests with the conflated module. In all cases it can be seen that the addition of G1 to the battery produced small gains in reliability that could have been achieved with additional items on the reading tests.

Table 9

Effect of G1 on Reliability of Objective Battery

	Reliability			
G1G2	**ALONE**	**COMBINED**	**N**	**ITEMS**
M1	906	924	177	117
M2	909	935	79	117
M3	857	919	88	111
M4	933	949	240	117
M5	977	964	41	122

A second omission from the final test battery was the conflated version of the test. The fifth module was constructed as a combination of the academic modules and a separate set of specifications was to be developed for the module. Despite the administrative gains that were to be had by the development of a single academic module, the face validity of special purpose modules led the steering committee to omit the conflated module from the test battery as well.

Probably the most difficult issue to address is the reliability of productive skills in language. Constable and Andrich (1984) examined the circumstances in which judges are required to assess productive type skills and are required to give ratings of performances. The usual case in which raters are trained to give similar ratings were examined and the paradox of higher correlations among the performances with constancy of ratings among raters, leading to higher reliability and lower validity were discussed. The recommendation of application of person judgement interaction was recommended and is followed in this examination of reliability of the writing scales.

Traditional notions of reliability depend on the degree to which the method of assigning scores eliminates measurement error. Four potential sources of error have been identified for the assessment of writing. These are..

(a) The writer within-subject individual differences,
(b) Variations in task
(c) Between-rater variations
(d) Within-rater variations.

To reduce within-subject error, a pool of similar tasks is often used. However, since essay writing is time consuming it is often logistically difficult to have students write several essays under examination conditions. In the IELT System the largest number of writing tasks set for any candidate is three in the General Training module. In all other modules the candidates are asked to write just two essays and there is a deliberate attempt to vary the nature of the task in order to increase the sample of writing styles. This is typical of essay examinations as task structures often differ with variation in topic. Within-subject task based variation has been traditionally difficult to control. In reducing variability due to task two parallel assignments or tasks have often been used. The most prevalent issue associated with writing assessment reliability is that of inter rater reliability. Statistical indices of agreement include coefficient alpha, generalisability coefficients, point biserial correlations, and simple percentages of agreement.

The most effective method found to reduce variation between raters is to provide training on specified criteria. Control of within-rater variability over time involves the use of periodic checks and common reference standards such as exemplar essays. However, in assessing raters as well as the ratings for reliability it may be useful to examine the stability of individual ratings and of tasks in terms of the

attribute being assessed.

The traditional definition of reliability from the classical or "true score" model is the proportion of variance that is due to the sample's true score variance. It depends on the average error variance of the test which arises from a variety of sources. Reliability is often estimated by calculating the correlation between repeated measures of the same entity such as an essay. However, reliability is a property of a variable not the test. It is a property f the measure that is obtained from the test. This can be interpreted as a line along which objects (on this case essays) can be positioned. The positions on the line need to be interpreted so equally spread intervals are required. In the assessment of language these are usually defined by various descriptions of language behaviour which are placed on a rating scale. In this case the rating points on the scale form the levels of proficiency used for reporting the assessments. Often the rating points are assumed or declared rather than defined via empirical methods. One empirical method for calibrating the units of measurement on the variable is through the application of item response theory (IRT). This brings together the notion of a person ability (or judgement) and the quality of an item (or essay) and enables a probabilistic statement about the person's judgement and the essay quality.

The rating assigned to an essay by a judge depends on a number of things. It depends on the quality of the essay and the dimension of quality that the judge uses. In proficiency assessment, the judge would be expected to use accepted notions of proficiency to assess the student as exhibited in the sample of writing in the essay. It depends on the raters ability to interpret the writing proficiency. This could be called the rating tendency of the rater and is commonly called the "rater effect".

It is typical of language assessment that the same set of rating points is used by all judges with every essay. Because of this it is usually considered that the relative proficiency levels associated with the rating points should not vary from essay to essay. That is an interpretation of the score of level 1 remains constant as do the interpretations of each level on the band scales. This consistency of score interpretation is usually associated with a fixed scale in this case called the band scale. For this reason a Rasch rating scale model has been adopted (Rasch, 1960; Andrich, 1978; Wright and Masters, 1982)

The model is defined by the equation:

$$P_{nix} = \frac{e^{\sum_{j=0}^{x}(\beta_v-(\delta_i+\tau_j))}}{\sum_{k=0}^{m} e^{\sum_{j=0}^{k}(\beta_v-(\delta_i+\tau_j))}}$$

where P is the probability of a specific rating being assigned,
x = 1 to m represents the number of steps in the rating scale.
T is the half distance on the variable between rating points and is therefore the threshold from one rating point to another.
d is the proficiency level for a specific rating point.
B is the rater tendency of the judge.
j is the number of essays judged over the m steps.

In this model successive levels are "recognised" once a threshold is passed so that d is the essay competence level and d +T is the threshold at which the judgement changes from a 1 to a 2.

The latent trait or variable is defined b the performances on tasks which require increasing amounts of attribute or proficiency. In this case however, the tasks are set and the performances vary according to student proficiency variation. If the trait exists among the judges, than they would sort essays according to their perception of the amount of trait exhibited in the essays and "levels" along the variable. Sorting would be based on the amount of writing proficiency. So the group of "expert" judges were asked to sort essays. If the sort of essay scripts were consistent across judges then a recognisable variable will have been identified and rater reliability should be high. If the sort were inconsistent across essays and

individual essays were assigned to too great a range of levels the reliability would be low and no underlying variable could be identified.

With consistent sorting the criteria used by the judges can be used to define the nature of the variable and would ultimately define the criterion scale. It is possible that the same set of essays could be sorted according to a range of criteria, each defining a different underlying variable. Where this is the case, the "sort" might be erratic and individual essays would not be consistently assigned to levels. Moreover the judges would not be consistently ordered with respect to their "rater effect". Under these circumstances, the reliability of the variable and the reliability if the judges would be low.

With these principles of item response theory in mind a series of workshops were organised in which judges would sort essays, articulate their criteria and establish a basis for estimates of both inter and intra rater reliability. However, the usual approaches to reliability estimation developed through classical item analysis are inappropriate and tend to give false information about the definition of the variable and the fit of the judges and the essays to the variable. Skehan's (1988;1989) papers point out the advantages of the item response theory approach to reliability estimation. However, there is an added advantage to those listed by Skeehan in that generalisability theory can also be used arising from the use of item response theory.

In assessing writing competence, essays are used as samples of work and a homogeneous set of essays can be used to define the rating points representing levels or levels on a variable defined as "writing competence". This is the first step in investigating the average variation in marking and identifying the components due to true score, the extent to which the essays do actually define a variable of writing competence and the extent to which raters use specified criteria. Two pieces of information then become available. Each essay can be assessed for its deviation from an expected position on the variable and its "fit " to the variable together with the estimates of error used as an estimate of its reliability. That is, reliability can focus on the essay at an individual level, and at the individual candidate level.

Given that essays are used to define the variable, the raters can also be placed along the variable using item response theory according to their predisposition for marking high or low on the variable (or placing essays in relative locations on the variable). If the variable is also defined for the raters in terms of specified criteria or descriptions of writing competence, than the variability among raters can be specified in terms of those descriptions. The information obtained from these procedures and the latent trait analysis may enable an examination of issues related to the effect of moderation, training and exemplar scripts.

Table 10.
RATER STATISTICS

	ASSESSMENT 1			ASSESSMENT 2		
NAME	MEAS	ERROR	FIT	MEAS	ERROR	FIT
A	.12	.34	-2.73	.16	.22	-4.28
B	.61	.36	-2.84	.32	.23	-4.39
C	.32	.31	-2.27	.20	.19	-3.24
D	.53	.28	-1.73	-.19	.23	-4.36
E	.53	.34	-2.63	.25	.20	-3.49
F	.22	.29	-1.84	.17	.15	-1.02
G	.12	.27	-1.59	.59	.18	-2.84
H	.64	.21	.27	.73	.15	-1.13
I	.36	.37	-2.97	.53	.20	-3.34
J	.53	.30	-2.06	.41	.24	-4.69
K	.19	.26	-1.38	.43	.30	-6.15
L	.53	.25	-1.03	.62	.15	-1.10

For assessment 1, six of the twelve raters appear to "fit" the underlying variable. On occasion 2 however, few raters appear to "fit" the variable. There appears to have been a change in the criteria or in the nature of the variable being used to assign scripts to levels. The original criteria used in the familiarisation workshop and reinforced in the training workshop, do not seem to have been used for assessment 2. Unfortunately it was assumed that the criteria would remain the same, and were in fact supplied to the raters. One curious point to examine is whether the apparent change in the criteria being used alters the rank order of the scripts for assessment two. This is examined in the analysis of script levels presented in Table 11 below. The results suggest that the rank order on the variable and the way in which the scripts have been assigned has not changed enough to warrant the rejection of the assigned scores. There is clearly a problem with the scoring of scripts in that the raters do not use a common set of criteria, neither when engaged in moderation nor when scoring solo. The training and selection of markers and their stability f ratings have become a focus of the IELTS management and scorers are required to undergo training with regular updates and monitoring to ensure that there is consistency among those chosen and those retained. These issues identified in the trials have assisted in developing appropriate training and monitoring procedures to ensure the consistency of raters used in marking the essay scripts.

Even in these trials, from a training perspective, there is a noticeable reduction in the variation of rating tendency, but the cost is high in terms of the ability of the raters to place the scripts along the variable of increasing competence. While the analysis appears to highlight this weakness, it would not be apparent under normal or classical analyses. The factor introduced to the assessment was the use of reference scripts and a consensus approach to allocation to levels. As can be seen later, the allocation to levels was not unanimous and three raters whose scores differed by considerable amounts adhered to their judgements leading to large residuals in the analysis, lack of fit among the raters and for the reference scripts. Despite this, there appears to be a maintenance of the range of script scores a move towards the ideal effect of training. That is, the range of ratings for the scripts has been maintained, covering ratings from 3 to 9 but the range of rating tendencies has been diminished. However, the analysis points out the problem of achieving this. There must be changes in the intra-rater scores in order to get this result. Hence there has been a loss of inter rater reliability from the first to second and third rating occasions. Moreover, the high agreement among raters on the second and third ratings means that there is very little variance and hence classical reliability estimates will be low. This is in fact the case, as the Latent trait estimates of person separation indices are low for occasion 2 and 3. (0.40 and 0.39 respectively). The item separation indices are high however, at levels of 0.74 and 0.77. (Wright and Masters, 1982). These indices reflect the discussion of Figure 1 and indicate the dilemma of rater studies. Low separation of raters needs to be coupled with higher separation of scripts. Hence the item response analysis in rater studies needs a very low person separation index and a high item separation index. These results appear to suggest that even after training, raters revert to their own criteria when marking solo. The implications for method of marking appear obvious. Moderation of non clerical marking procedures is essential.

While the raters did not appear to be consistent with the application of criteria, the effect on the bands did not seem to be as severe.

Table 11
Script Assessment Time 1 and Time 2.

NAME	T1	ERROR	FIT	T2	ERROR	FIT	T2-T1
M11A	3.23	.35	-1.92	3.38	.26	-1.11	0.32
M11B	4.33	.35	-.80	4.26	.28	.19	0.10
M11C	1.96	.30	-.74	1.62	.41	-1.61	-0.17
M11D	2.62	.31	-1.45	2.64	.27	-.83	0.19
M11E	1.69	.42	-1.83	1.74	.43	-1.86	0.22
M12A	2.88	.32	-1.64	3.12	.35	-2.52	0.41
M12B	3.18	.26	-.67	2.68	.29	-1.14	-0.33
M12C	2.07	.33	-1.32	2.01	.35	-1.25	0.11
M12D	**2.84**	**.41**	**-2.70**	**3.25**	**.73**	**-5.29**	**0.58**
M12E	1.79	.42	-1.77	1.68	.37	-1.28	0.06
M21A	2.36	.42	-2.47	2.36	.35	-1.56	0.17
M21B	3.86	.30	.99	3.76	.25	-.40	0.07
M21C	1.36	.33	-.70	1.16	.32	-.35	-0.03
M21D	2.40	.41	-2.46	2.07	.30	-.58	-0.16
M21E	2.62	.53	-3.52	2.50	.36	-1.83	0.05
M22A	2.54	.34	-1.79	2.59	.54	-3.38	0.22
M22B	3.49	.36	-1.75	3.76	.51	-3.64	0.43
M31E	1.56	.28	.32	1.50	.29	-.00	0.11
M43C	4.06	.32	1.23	4.78	1.09	-2.60	0.89

Shifts in the values assigned to scripts were examined using common item equating. methods. Mean item measures for each occasion were used to compute the link shift for the items. (0.17). In the table only adjusted "attribute "values are shown. Three scripts changed from "non fit " to "fit" on the second assessment and three scripts reversed this. All others in the link set were found to "fit' the writing proficiency variable. While the raters have unstable "fit" characteristics, this may have been influenced by the new scripts marked on the second occasion. It does not seem to have influenced the ranking of scripts from the initial assessment.

It is noticeable that scripts with high fit statistics also have the largest translation shifts associated with the equating across occasions (T1-T2). This indicates that these scripts have characteristics which tend to confuse the ratings and introduce secondary characteristics not included in the criterion scales. However, the size of the fit statistics is expected to be large, given that there were only 15 raters on each occasion and 43 scripts on occasion 1 and 20 scripts on occasion 2. The effects of training should be observable in the consistency of the ration as discussed above. Probably the most telling information is the change in the "fit" statistic. The test used is commonly called the Infit statistic, which applies a chi-squared-like test to residuals. The test is sensitive to outliers. Hence the effect of raters whose judgement differs considerably from others will have an enhanced effect. This is mostly the case with scripts

This study has illustrated that conventional estimates of rater reliability loose much of the available information and enter the researcher into a paradox when inter rater reliability is maximised. By reducing the variation among raters, the classical approach to reliability is jeopardised. Latent trait analyses provide item and person specific measures of reliability or error variance and these may be used to advantage in examining trends in the data.

References:

Andrich, D. Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 1978, 2, 581-594.

Andrich D. and Godfrey, J.R. Hierarchies in the skills of Davis' Reading Comprehension Test Form D: an empirical investigation using a latent trait model. *Reading Research Quarterly*. 14, 2, 182-200.

Constable, E. and Andrich, D. (1984). Inter Judge Re;liability: Is complete agreement among judges the ideal? Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, April 24-26.

Griffin, P. The development of the IELTS test battery. Paper presented at the annual conference of the Australian Applied Linguistics Association. Launceston, July, 1988.

Rasch, G. (1960, 1980) Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedegogiske Institut. and Chicago: University of Chicago Press.

Skehan, P. (1988) Peter Skehan on Testing. Part I. Language Teaching, 21,4, 211-221.

Skeehan, P. (1989) Peter Skehan on Language Testing. Part II. Language Teaching, 22, 1, 1-13.

Wright B. and Masters, G. (1982). Rating Scale Analysis. Chicago: MESA Press.

Wright B. and Stone, (1979) Best Test Design. MESA Press, Chicago.